Low Cost, Generic Molecular Markers for Breeding and Research

Dave Hoisington and Rajeev Varshney, ICRISAT, Patancheru, Andhra Pradesh, India
Andrzej Kilian, DArT, P/L, Canberra, ACT, Australia
E-mail: *d.hoisington@cgiar.org*

Cutting Edge

The use of molecular markers has been shown to be a powerful approach for the genetic dissection and manipulation of many traits of importance in agriculture. However, their full impact, especially in breeding, is yet to be realized. This is due to a number of limitations such as the time and resources required to discover a large number of polymorphic markers for a species, requirements for sequence information for marker development, and the cost and time for scoring the markers (average per datapoint cost of over US$1.00).

With the advent of Restriction Fragment Length Polymorphisms (RFLP), genetic maps were developed for many species, including many crops of agricultural importance. While various strategies were employed to reduce the costs and increase the throughput of RFLP detection, the lengthy and elaborate procedures of Southern blotting and probe hybridizations greatly limited the number of samples that could be analyzed, thus maintaining high per sample costs. The discovery of microsatellites (or simple sequence repeats, SSRs) in many species provided much simpler protocols based on PCR amplification and gel-or sequencer-based detection. However, there are large costs and time required to develop primers that detect SSRs in a species, although once developed, SSR assays are relatively simple, low-cost and can be automated for moderately high-throughput. An often overlooked limitation of SSRs is the lack of an efficient automated analysis system to detect the specific SSR allele in a sample, and to convert this into usable data for mapping or characterization (software does exist but still often requires significant user input for interpretation).

Recently, systems that detect single nucleotide differences between DNA samples (single nucleotide polymorphisms, SNPs) are being developed as these can be assembled into large-scale, low-cost detection platforms such as the Illumina Bead Arrays (www.illumina.com). For species where extensive sequence information exists (e.g., maize and rice), the development of SNPs is relatively easy, although overall costs can be high. For other species, the requirement for extensive sequencing is a limitation, although such sequencing is becoming more achievable with recent developments in extremely high-throughput sequencing strategies (e.g., Life Sciences 454 DNA Sequencing, www.454.com).

> **"DArT technology is not dependent on having DNA sequence information, is expandable based on user's needs, and is high-throughput with low cost"**

Diversity Arrays Technology (DarT)

A marker platform that offers certain advantages, especially in the breeding context, is based on array hybridization technology and has been termed, Diversity Arrays Technology (DArT). Developed by Dr. Andrzej Kilian and his team, this technology provides a cost-effective and whole-genome genotyping tool (see references and the DArT P/L web site: www.diversityarrays.com). The technology is not dependent on having DNA sequence information, is expandable based on the user's needs, and is high-throughput and low cost (per datapoint costs below US$0.10). DArT analysis requires the assembly of a "genotyping array" that contains genomic clones that show variable hybridization intensities across the individuals under study. Such arrays have been developed by Dr. Kilian and his team for several species, including sorghum, rice, barley, wheat, chickpea, pigeonpea and many others including animal species. The "genotyping array" is hybridized with an individual DNA sample using microarray technology, the hybridization signal measured and converted into a genotype score. DArT P/L bioinformaticians have developed specialized software for automated data extraction, storage and analysis.

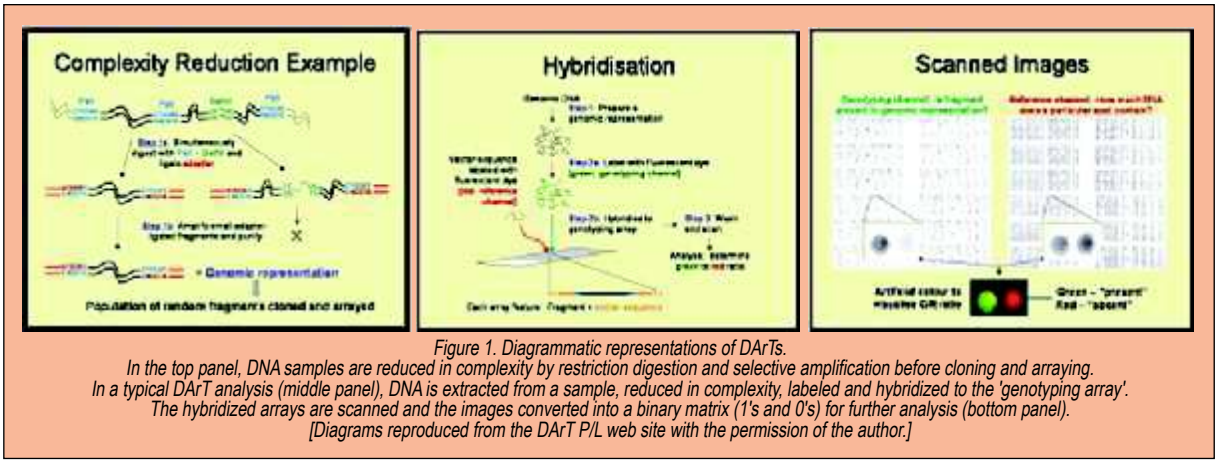

Figure 1. Diagrammatic representations of DArTs.
In the top panel, DNA samples are reduced in complexity by restriction digestion and selective amplification before cloning and arraying.
In a typical DArT analysis (middle panel), DNA is extracted from a sample, reduced in complexity, labeled and hybridized to the 'genotyping array'.
The hybridized arrays are scanned and the images converted into a binary matrix (1's and 0's) for further analysis (bottom panel).
[Diagrams reproduced from the DArT P/L web site with the permission of the author.]

Diversity Arrays Technology (DArTs) Low Cost, Generic Molecular Markers for Breeding and Research

DArT Development

A DArT marker is simply a segment of DNA that has been determined to be polymorphic across a range of germplasm of interest. Such markers are most easily scored as dominant (presence/absence) loci, although many can be scored in a hemi-dominant manner. Like all marker systems, a marker discovery phase is required to identify the most polymorphic and useful markers for large-scale genotyping. Unlike most other marker systems, no sequence information or DNA synthesis is required, and the same platform is used for both discovery and routine genotyping.

Figure 1. *(on page 6)* presents the basic steps in a DArT analysis. To develop a genotyping array for a species, genomic DNA is isolated from a number of individuals that represent the diversity to be studied (e.g., genebank accessions, breeding lines, parental lines). The pooled DNA is then reduced in complexity by one of several methods that involve digestion with one to three restriction enzymes of various cutting frequencies, followed by adapter ligation and PCR amplification. The 'representation' produced following amplification is then cloned and individual inserts (markers) are arrayed on a microarray. The production of markers via cloned DNA provides a rapid and simple method for generating the sequence of any specific DArT marker in the future. Labeled genomic DNA from the original individuals is then hybridized to the array and polymorphic clones (DArT markers) are identified that produce hybridization signal intensity differences for different individuals. The selected markers are then used to produce a 'genotyping array' for large-scale genotyping. Several thousand markers can be screened on a single array to discover the most polymorphic markers that are then used on 'genotyping arrays' to detect several hundred loci in parallel using only small quantities (ng) of sample DNA.

For routine DArT genotyping, sample DNA is reduced in complexity using a similar protocol used in the discovery phase, labeled, and hybridized to one or more genotyping arrays. The hybridization signals are measured and specialized software (DArTsoft) used to convert the signals into genotypic scores (usually 1's and 0's). The genotypic scores can then be analyzed in a manner similar to other molecular marker genotypes (e.g., SSRs, SNPs) using statistical software packages for diversity, genetic and/or breeding applications.

The infrastructure required for high-throughput DArT analyses is not extensive, but does require a dedicated molecular biology laboratory with certain basic equipment for DNA extraction, cloning and microarray hybridization, along with more specialized equipment for microarray printing and scanning (Figure 2. *on page 8*). In addition, adequate computing support is required to handle the large datasets that are produced.

DArT Applications

Dr. Kilian and his team at DArT P/L have made excellent progress in applying DArTs in a number of plant, animal and pathogen species. For plants, these include important agricultural species such as rice, barley, wheat and sorghum, among other important crop and model species. Applications of DArTs are essentially the same as for other marker types. DArT arrays can be extremely useful in assessing the diversity of genetic resource collections and breeding germplasm. The ability to genotype an individual at relatively low per datapoint costs means that a much larger set of individuals (accessions) can be studied, thus often eliminating the need to identify subsets of the materials due to cost and technology constraints. By analyzing the entire collection, a complete understanding of the diversity is obtained, greatly reducing the risk of missing critical germplasm accessions in the study.

The generation of molecular maps that form the basis to locate genomic regions (Quantitative Trait Loci, QTLs) involved in the expression of important agronomic traits has been and will remain an important application of molecular technology. To obtain even a reasonable level of genome coverage and a fairly complete genetic map of a species, requires a few to several hundred molecular markers. In some species (e.g., rice, maize, Arabidopsis), markers are abundant and polymorphic so that the identification of sufficient quantity of molecular loci is easy. In most other species, there is still a lack of sufficient numbers of polymorphic loci. While DArT markers may not be any more polymorphic than other marker types, the fact that several thousand can be screened very rapidly means that DArTs provide an efficient marker type for whole genome mapping. Once a few hundred DArT markers are available for a species, the development of a molecular genetic map for a segregating population becomes a matter of days, rather than months or years that other marker types can require. In fact, by using DArTs, the major limitations in the genetic analysis of most traits are the availability of segregating populations and/or cost-effective phenotyping protocols.

Association mapping is becoming an option in many species, especially for those where the development of genetic populations is restricted or time consuming. Effective association mapping requires fairly dense coverage of a genome with molecular markers. Again, the large number of DArTs that can be screened in a single assay provides an attractive option for such studies. Once particular associations are identified, the same DArT array can then be used in backcrossing the associated segments into elite breeding and farmer varieties.▶▶

"The ability to genotype an individual at relatively low per datapoint costs means that a much larger set of individuals (accessions) can be studied"

Diversity Arrays Technology (DArTs) Low Cost, Generic Molecular Markers for Breeding and Research

Of particular importance is the application of DArTs to breeding programs, especially for use in marker-assisted selection (MAS). A major limitation for the use of markers in breeding is often the lack of a large number of markers, both linked to the trait(s) of interest, and for selection of the background (or non-linked) genomic segments (e.g., in a backcrossing program). Beyond just the number of markers, the cost of, and time needed to, produce a datapoint is often prohibitive. Given the extensive genome coverage, high-throughput and low-cost, DArTs provide a very attractive marker system for such breeding applications. Costs of below US$0.10 per datapoint mean that a single individual in a backcrossing program could be genotyped for several QTL segments (using two flanking markers) for around US$1. By including 100 additional DArT markers on the array to detect random loci throughout the remainder of the genome, an entire genotype of the backcross individual would be determined for around US$10.00. Such complete genotypic information would allow the breeding program to select those individuals that not only have the QTLs of interest, but also contain the maximal amount of recurrent parent genome greatly accelerating the conversion of elite lines. All of this selection can be accomplish well before the breeder makes the actual crosses in the field, e.g., at the seed and/or seedling stage.

"Given the extensive genome coverage, high-throughput and low-cost, DArTs provide a very attractive marker system for breeding applications"

A simple application of DArTs for MAS is in the backcrossing of individual or multiple transgenes into new varieties. Given the difficulty to produce a highly effective transgenic event and the fact that most regulatory policies consider each 'event' as unique, once an 'event' has been developed that will be released in a country, it will be necessary to introgress this 'event' into a range of elite varieties. This is accomplished in most crops by backcrossing for several (4+) generations followed by selfing for 2-3 generations to fix the transgene in the final product. At each generation, the presence of the transgene is determined by various methods ranging from phenotpying, to ELISA, to DNA detection. While such methods are simple and inexpensive, most backcross programs are not able to make any selection on the recurrent parent genome as this would require extensive genome analysis at a number of loci. A single DArT genotyping array with ~100 markers would provide a low-cost method to determine how similar a particular backcross individual is to the desired recurrent parent. This could result in obtaining the desired new variety in 50% or less of the time required if DArTs were not used.

Given the advantages offered by DArTs and the desire to provide a generic technology that could be applied to any species of interest, The Center of Excellence on Genomics (CEG) recently approved by the Department of Biotechnology (DBT) to be established at ICRISAT, will implement DArT technology as one of its flagship marker services for research and breeding in India. A partnership with DArT P/L in Australia will provide the required technology transfer and training to ensure that DArTs are effectively established and operating systems provided in the most timely fashion. Initially, DArT systems will be available for a number of important crop species in India. Other species will be added through collaborations with Indian institutes and scientists interested in using DArTs in their research.

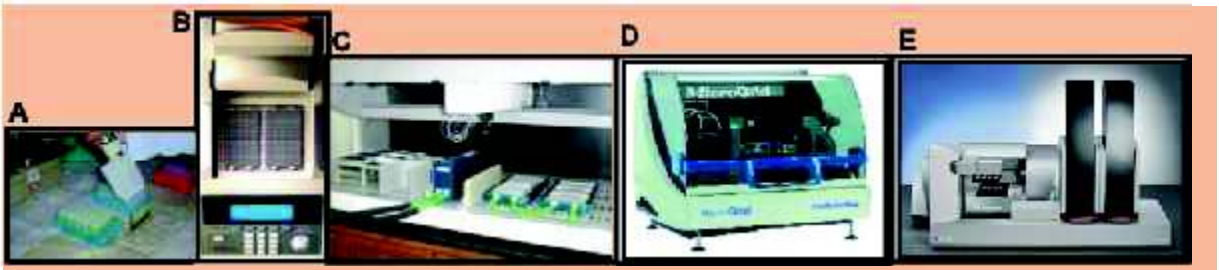

Figure 2. Examples of the major infrastructure required for DArT analyses. In addition to a basic molecular biology laboratory equipped for DNA extraction (A) and PCR amplifications (B), the use of a robotic pipetting station (C) for more automated DNA extractions and PCR setup can be useful. The two major pieces of equipment required are a microarrayer (D) to print the arrays contained the DArT markers and a scanner (E) to detect the signals following hybridization with sample DNA.

While the CEG will be providing DArTs as a major marker service, other technologies such as DNA sequencer-based SSR detection, TILLING and EcoTILLING detection via LiCOR systems, and high-capacity phenotypic analysis for abiotic stresses will also be available at the CEG. High-capacity data storage and analysis will be established to provide the necessary data handling for the large datasets produced by the CEG and to ensure that these are made available to the researchers in a timely and effective manner. The ultimate goal of the CEG is to make molecular markers available as an effective tool for all researchers and breeders in India.

(More details and references are available on the DArT P/L website, www.diversityarrays.com